Portland State University
PDXScholar

Computer Science Faculty Publications and Presentations

Computer Science

6-2018

Challenges and Opportunities in Transportation Data

Kristin A. Tufte
Portland State University, tufte@pdx.edu

Kushal Datta
Intel

Alekh Jindal
Microsoft

David Maier
Portland State University, maier@cs.pdx.edu

Robert L. Bertini
Portland State University

Let us know how access to this document benefits you.

Follow this and additional works at: https://pdxscholar.library.pdx.edu/compsci_fac


Part of the Transportation Commons

Citation Details

Tufte, K., Datta, K., Jindal, A., Maier, D., & Bertini, R. L. (2018, June). Challenges and Opportunities in Transportation Data. In Proceedings of the 1st ACM/EIGSCC Symposium on Smart Cities and Communities (p. 2). ACM.

This Conference Proceeding is brought to you for free and open access. It has been accepted for inclusion in Computer Science Faculty Publications and Presentations by an authorized administrator of PDXScholar. For more information, please contact pdxscholar@pdx.edu.

Challenges and Opportunities in Transportation Data

Kristin Tufte
Portland State University
tufte@pdx.edu

Kushal Datta
Intel
kushal.datta@intel.com

Alekh Jindal
Microsoft
aljindal@microsoft.com

David Maier
Portland State University
maier@cs.pdx.edu

Robert L. Bertini
University of South Florida
rbertini@usf.edu

ABSTRACT

From the time and money lost sitting in congestion and waiting for traffic signals to change, to the many people injured and killed in traffic crashes each year, to the emissions and energy consumption from our vehicles, the effects of transportation on our daily lives are immense. A wealth of transportation data is available to help address these problems; from data from sensors installed to monitor and operate the roadways and traffic signals to data from cell phone apps and — just over the horizon — data from connected vehicles and infrastructure. However, this wealth of data has yet to be effectively leveraged, thus providing opportunities in areas such as improving traffic safety, reducing congestion, improving traffic signal timing, personalizing routing, coordinating across transportation agencies and more. This paper presents opportunities and challenges in applying data management technology to the transportation domain.

KEYWORDS

Data Management, Smart Cities, Transportation Data

ACM Reference Format:
Kristin Tufte, Kushal Datta, Alekh Jindal, David Maier, and Robert L. Bertini. 2018. Challenges and Opportunities in Transportation Data. In *Proceedings of The 1st ACM/EIGSCC Symposium On Smart Cities and Communities (SCC2018)*. ACM, New York, NY, USA, 8 pages.

1 INTRODUCTION

Through innovations in transportation, our world has become increasingly connected. However, transportation comes with multiple challenges. It is estimated that in 2016, 37,461 people in the United States died in fatal car crashes [34]. This number represents a 5.6% increase in fatalities over 2015 and the highest number of traffic fatalities since 2008 [34]. Fatalities are not the only cost of transportation. In 2014, commuters spent 6.9 billion hours in congestion, wasting 3.1 billion gallons of fuel and costing the U.S. economy an estimated $160 billion [50]. Finally, transportation accounts for 27% of greenhouse gases produced in the United States [13]. With increasing vehicle travel in the U.S. [54] and the dramatic increase in privately owned cars in developing countries, these adverse effects of transportation systems are felt across the globe.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
SCC2018, June 2018, Portland, Oregon USA
© 2018 Copyright held by the owner/author(s).
DOI: 10.1145/3236461.3241971

Interestingly, transportation systems are relatively well instrumented; sensors and surveillance are commonly installed on freeways, on buses and trains, at traffic signals, in bicycle lanes and more. In addition to such fixed sensors, vehicles are well instrumented, and private-sector companies gather transportation data from location data from cell phones and vehicles [20, 23, 30, 47]. Finally, the federal government, private automakers and after-market developers are rapidly developing automated and connected vehicle technology [9]. This technology is expected to generate vast amounts of data, in excess of 20 petabytes/second by some estimates [53]. Transportation data is regularly gathered and archived [6, 10, 15, 24, 39, 41], though approaches are inconsistent. Better use of transportation data and archives could improve real-time traffic signal timing; personalize routing; improve reactions to traffic crashes by coordinating response across different transportation agencies; and improve understanding of traffic congestion and bottlenecks. All of these would help improve safety, reduce congestion and limit carbon emissions. There is a wealth of data available from transportation systems that has the potential to help address several critical issues and therein lies opportunity.

Leveraging transportation data for applications such as safety improvements and congestion and emissions reductions requires a variety of new data management technologies. New data models such as multi-graphs or semantic ontologies may be required to represent the complex relationships among different transportation modes. New data integration techniques are needed to combine the truly varied, multi-source transportation data in real time. Existing physical system designs do not effectively fit the wide variety of dynamic transportation data that needs to be analyzed. New query processing techniques are needed to process richer declarative queries which arbitrarily select and combine different pieces of the transportation graph. The dynamically changing nature of transportation data requires new update and query semantics. And finally, an understanding of how to systematically re-use data collected for one purpose (operating the transportation system) for another purpose (research or planning) is required. This paper contributes a description of opportunities and challenges in the transportation domain based on 20 years' work in this arena.

The remainder of the paper is organized as follows. Section 2, discusses trends in and characteristics of transportation data. In Sections 3 and 4, we describe different personas and use cases in the transportation domain. Finally Section 5 considers open data management research questions in the transportation domain.

2 THE TRANSPORTATION DATA DOMAIN

Transportation data is diverse. It includes data from fixed sensors with relatively low volumes and velocity, crowd-sourced data from apps such as Google Maps [17] and Waze [56], and data from connected and automated vehicles with potentially very high volumes and velocities. Transportation data is regularly archived; those archives provide a rich data source that can be analyzed to help solve societal problems. In this section, we survey transportation data. We begin by discussing newer data sources such as connected and automated vehicle data, and probe-vehicle and crowd-sourced data. We then discuss transportation data archives and traditional data sources and conclude with a brief analysis of data volumes and trends.

2.1 Connected and Automated Vehicle Data

Perhaps the most exciting new transportation data source is connected vehicles. At a high level, connected vehicles are vehicles that communicate with other vehicles using “Vehicle-to-Vehicle" (V2V) technology or to the infrastructure using “Vehicle-to-Infrastructure" (V2I) technology. The federal government is pursuing Connected Vehicle programs based on Dedicated Short Range Communications (DSRC) [25] along with auto manufacturers and other partners; SAE J2735 [42] is the likely message set for connected-vehicle communications. The connected vehicle data sets have enormous possibilities — both in terms of the potential to be very large (10-27 petabytes/second by some estimates [53]) and very useful.

In addition, private automakers are increasingly automating their vehicles with functions including adaptive cruise control, lane keeping, automatic braking for collision avoidance, steering and parking to name a few. Recent developments include tests of fully-automated vehicles, including trucks [18, 48].

Advances in connectivity and automation promise improved and safer driving experiences as well as a wealth of data. Today’s vehicles already internally amass large amounts of data, including location, speed, acceleration, brake status, windshield-wiper status and temperature [7]. According to the US Department of Transportation (DOT) connected vehicles can improve mobility, safety and the environment [25, 55]. Samples of transportation data, including connected vehicle data, are publicly available through the USDOT ITS Public Data Hub [22].

2.2 Probe Data and Crowd-Sourcing

Another source of transportation data comes from probe vehicles — vehicle position data and other attributes often collected through cell phone apps or other mechanisms. When a user allows an app such as Google Maps to use the location of their phone, location information is anonymously sent back to Google [16]. This “crowd-sourced" data is used by applications such as Google Maps to estimate speed and provide traffic information, but may also be useful for transportation planning and other purposes. The National Performance Management Research Data Set (NPMRDS) [35] is another example of probe data. The NPMRDS data can be up to 1-3 GB per month for larger states [29]. A final example of crowd-sourced data is Waze (now part of Google) [56]. Beyond standard passive crowd-sourcing of location and traffic information, Waze further allows users to actively enter information about events such as crashes; this information is then incorporated into the Waze interface and can also be provided to public agencies through the Waze Connected Citizens program [56].

2.3 Fare-Card Data

Transit systems are beginning to use RFID smart cards and other connected apps for transit payment; the ORCA card in Seattle, WA [37] and the Hop Fastpass [14] in Portland, OR are two examples. Data from those cards — such as where people get on and off transit — is collected and is potentially very valuable to researchers and transportation planners as such data can be used to understand people’s travel and trip patterns.

2.4 Transportation Data Archives

The transportation community has realized the value of archiving data collected for the purpose of operating the transportation system, recognizing that this data is also useful for other purposes. For example, data gathered while operating traffic signals — turning the lights green, yellow and red — if archived and stored, can be used to improve signal timing. Archives that gather and store such operational data are becoming commonplace and include PORTAL [39], iPEMS [24], RITIS [41] and DriveNet [10]. Cities such as Atlanta [6] and Dublin [15] are also developing archives.

To illustrate, PORTAL is the official transportation data archive for the Portland, OR – Vancouver, WA metropolitan region [39, 51]. PORTAL archives over ten types of data (most are live data feeds) from six different transportation agencies in the region into a ~3TB PostgreSQL database. Major types of data in PORTAL include freeway traffic speeds and vehicle volumes, arterial traffic volumes,

Data Type	Agencies
Freeway Speed, Volume, Occupancy	ODOT, WSDOT
Freeway & Arterial Travel Times	ODOT, City of Portland, Clark County, Washington County
VMS & VAS Sign Data	ODOT
TOC Incident Data	ODOT
Arterial Volume & Speed	City of Portland, Regional, Clark County (Speed is Clark County only)
Central Traffic Signal System Data	City of Portland, Clark County
Passenger Counts, On-time Performance from AVL/APC System	TriMet, C-TRAN
GTFS Schedule Information	TriMet, C-TRAN
Weigh-In-Motion Data	ODOT
Bicycle Counts	City of Portland
Vehicle length data	ODOT, Washington County
Signal Phase and Timing Data (*Under Development*)	Clark County, City of Portland

Table 1: Data sources

arterial traffic signal data and travel times, transit (e.g. bus, light rail) data and freight data. (Note that "vehicle volume" is a flow measure, based on vehicle counts past a point over time). PORTAL has a web interface that provides analyses, customizable visualizations and data downloads at user-specified aggregation levels. Table 1 lists current PORTAL data sources, with data sources under development marked and Figure 1 shows a screenshot of the current travel time page on the PORTAL web site. Finally, a documented, sample PORTAL data set is available [38].

2.5 Freight: WIM and Length

Freight traffic accounts for 7% of traffic, but 17% ($28 billion) of the estimated $160 billion cost of transportation to the 2014 U.S. economy [50]. Understanding freight traffic is therefore important. Freight data comes in multiple types including "weigh-in-motion (WIM)" data in which trucks are weighed and their lengths and heights measured as they travel along key state highways for efficient enforcement of vehicle size and weight limits [32]. WIM data provides accurate vehicle-type (semi-truck vs. pickup-truck, etc.) and length data. Data on vehicle lengths is also available from standard freeway detectors (dual inductive loop).

Key Data Features: Structural and Quality Differences. The two types of freight data are similar in that both can be used to estimate the percentage of freight trucks on the road. The WIM data is detailed and accurate, but is available for limited locations, while the newer vehicle-length data is less detailed, but is available from many more locations.

2.6 Transit: Ridership and On-time Performance

Transit (e.g., bus, streetcar, and light rail) systems collect large amounts of data from Automatic Vehicle Location (AVL) and Automatic Passenger Counter (APC) systems on their vehicles. This data includes information such as whether a bus stopped at a stop, how long the bus was stopped, the number of passengers who boarded the bus, the number who alighted, and whether the handicap lift was used. Used by transit agencies for fleet management, the data are collected every time a bus stops at or passes a bus stop.

Key Data Features: High-Volume, Standardized Formats. Transit data is relatively high-volume. In addition, transit-schedule data and actual arrival times and locations are typically provided in a standard format called General Transit Feed Specification (GTFS) [19]; a format originally developed by a collaboration between TriMet, the Portland, OR transit agency [49], and Google [17], which is now used by more than 1,350 agencies.

Figure 1: Screenshot of PORTAL Travel Time Interface

2.7 Arterial: Vehicle Volume, Travel Time, Traffic Signal Phase and Timing

Loosely, arterials are major roadways with full access (not limited-access freeways) Arterials also typically have interrupted flow (not uninterrupted as on freeways) involving traffic signals or roundabouts for intersection control. Traffic is intended to stop on arterials (not on freeways).

Vehicle Volume: Arterial vehicle counts are collected from either high-definition radar or inductive loops and are typically reported at 1–5 minute intervals.

Travel Time: As arterials are interrupted-flow facilities (due to traffic signals), travel times along arterials are measured directly and are not interpolated from spot speed measurements. Vehicle arrival information at one location is matched with arrival information from a second location to determine travel time between those location. This matching can be done using technologies such as Bluetooth readers or license plate readers.

Traffic Signal Data: Sensors are installed in arterial roadways to help operate traffic signals. Arterial traffic signal data includes logs of traffic signal cycles, vehicle volume counts and signal phase and timing data — data detailing when traffic signals change phase.

Key Data Features: Point-based vs. Segment-based. A key feature of this data is that while arterial volumes are point-based by definition, arterial travel time is segment-based, requiring the ability to combine point-based and segment-based data.

2.8 Freeway: Speeds and Vehicle Volumes

PORTAL has archived freeway traffic speeds and vehicle volumes from the Oregon Department of Transportation (ODOT) since 2004 and from the Washington Department of Transportation (WSDOT) for Vancouver, WA since 2012.

Key Data Features: Multiple Sensing Types; Low Data Volume. Lane-level traffic speed and volume data is collected from multiple sensor types including traditional inductive loops and newer high-definition radar detectors which have different sensing characteristics. The differing characteristics do not impact relatively simple data usages such as creation of speed maps and calculations of travel times; however, the differences may impact more sophisticated analyses such as automatic identification of the location of freeway bottlenecks.

2.9 Data Volume and Trends

The volume and types of transportation data available have grown significantly in recent years and this growth is expected to continue.

Data growth has come from new data sources such as probe vehicle data, crowd sourcing and soon-to-be-available connected-vehicle data. Connected vehicle data may be tens of petabytes a second [53], Signal Phase And Timing (SPAT) is estimated at a TB/year or more for a small community and finally data collected from mobile phones by companies such as Google and Nokia with detailed travel data is expected to be multiple TB/s per year [26]. These data sets are all relatively large in volume and velocity, experience shows that the major impediments to using the data are issues including institutional barriers, modeling, combining, cleaning and repurposing the data. This section has summarized the rich transportation dataset, the next section describes the personas of the data users who may leverage this data.

3 PERSONAS

The transportation system and transportation data have a wide variety of stakeholders; as such transportation applications must serve a variety of users with varying needs. In this section, we describe four personas that use transportation data: Traveling Public, System Operators, Transportation Planners and Private-Sector Businesses.

3.1 Traveling Public

The traveling public use routing services, such as provided by Google Maps [17] or a transit agency, on-line speed maps, and variable message signs that provide travel times or warn of congestion ahead. In the not-too-distant future, we expect the transportation system to interact directly with on-board systems on vehicles (vehicle-to-infrastructure) and for vehicles to communicate with each other (vehicle-to-vehicle). Applications that customize information to users' preferences or combine data from multiple transportation modes will be important.

3.2 System Operators

System Operators are responsible for day-to-day operation of the transportation system. Operators make real-time decisions such as deploying incident response teams, adjusting traffic-signal timings or adding additional transit vehicles in response to traffic conditions or incidents. Operational systems are currently siloed with transit agencies managing and responding to transit data, departments of transportation managing freeways and responding to data from freeways, cities and counties managing arterials and local roads, and so on. Combining data across systems and agencies would be of great value to transportation departments. System Operators and the systems they manage have both hard and soft response-time limits and need reliable systems with high availability.

3.3 Transportation Planners

Transportation planners are responsible for long-term planning and decision making, such as evaluating options for reducing congestion, determining if a new lane is needed, fixing speed limits, supporting budgetary allocation priorities, and making transportation plans which require accurate estimates of current traffic levels. They also are responsible for calculating performance metrics such as service quality and sustainability. Their use of the system is offline, but includes integration and analysis of large volumes of diverse data.

3.4 Private-Sector Businesses

Private-Sector Businesses are rapidly getting into the transportation sector including companies such as Google, which provides mapping services [17] and is incubating Sidewalk Labs [45], Waze (now owned by Google), which allows the public to report traffic incidents [56] and Transportation Network Companies (TNCs), such as Uber [52] and Lyft [33], that connect passengers with drivers to provide ride services. RideScout is a relatively new application which allows a user to compare transportation options across modes (buses, bikes, TNCs, etc.) to give users choices [40].

4 POTENTIAL APPLICATIONS

Transportation data has a wide variety of applications. Selected potential applications are described in this section. A detailed list of connected vehicle applications can be found on the US Department of Transportation web site [25].

4.1 Multi-modal Routing

Traditional routing services, such as in Google Maps, provide the Traveling Public point-to-point routes between two locations, typically using a single mode of transportation and optimizing for shortest time or shortest distance. Routing services that integrate multiple modes in one trip, such as *drive to the light rail station, take light rail into the city, rent a bike from bikeshare to get to final destination* are desired and are under development in some cities. Routing based on user preferences such as *find the shortest route to my work stopping by the coffee shop I like* or more complex *find me most sustainable route to my home* are also desirable. Routing based on current or forecasted actual traffic conditions is also sometime available and very desirable.

4.2 Improved Transit Arrival Predictions

Today transit arrival feeds have information about when buses and trains are estimated to arrive at locations. Real-time transit arrival data (available in GTFS format [19]) feeds the "transit tracker" signs that are seen at bus or light rail stops or through apps. However, this data may not correctly predict future arrivals in part because it is based on current location, but does not integrate live information on road congestion. If live congestion data can be jointly analyzed with live real time GTFS data, the accuracy of arrival predictions can be improved. Combining these data sources requires on-line data integration and dynamic merging of geo-spatial data. Such accurate arrival data can also feed apps [3] that provide transit-arrival information.

4.3 Real-time Congestion Detection

In transportation, a bottleneck is a point on a roadway where congestion (queueing) occurs on a regular basis. For example, a bottleneck may be caused by a sharp turn or a narrowing of the roadway. It is common in transportation to use off-line analysis of traffic data and knowledge of the structural road network to identify bottlenecks. However, dynamic identification of congestion — recognizing congestion that occurs due to non-recurring events such

as inclement weather, maintenance, construction, crashes or special events — and its secondary effects where congestion on a primary road or highway causes congestion on alternate routes is also important. Traffic managers would like to identify and react to such secondary congestion. Live traffic updates in conjunction with static road network data can be used to model primary and secondary effects of congestion. Betweenness centrality measures might be used to identify secondary effects of congestion; however, doing so would require updating standard algorithms to run incrementally over on-line data.

4.4 Integrated Corridor Management

Integrated Corridor Management (ICM) means managing movement of people through a "transportation corridor." For example, many people travel back and forth from downtown Portland to its eastern suburbs; this route is considered a transportation corridor. One can use many different ways to travel from downtown Portland to the eastern suburbs — freeway (I-84), arterial (US 26/Powell Blvd), light rail, multiple bus lines or even a bicycle. These facilities are managed by different agencies; ODOT manages I-84, City of Portland manages US 26/Powell Blvd and TriMet manages the light rail and bus lines. The desired goal is for Operators to manage these facilities in real-time in an integrated manner, focused less on vehicles and more on people.

For example, in the case of a crash on the freeway, traffic signals on the parallel arterial can be re-timed to accommodate the additional traffic that spills onto the arterials when the freeway is congested. Additional buses and light rail trains could be dispatched.

Off-line analysis can answer two key questions of interest to transportation planners: 1) How do travel times compare across these different “modes" of transportation? and 2) What is the “mode split" between the different “modes" (freeway, arterial, transit) of transportation?

4.5 Improved Traffic-Signal Timing

The average driver spends many hours a year stopped at traffic signals. Currently, most traffic signals change phase based either on fixed plans that take into account time of day and day of week or on sensors local to the intersection, which indicate, for example, if there is traffic demand on a side street. Such systems do not take into account the general state of the system. Further, for the signals running on “fixed plans," traffic-signal timings are updated every three to five years depending on signal location and municipality.

The SCATS System [43] uses real-time data from in-road sensors to dynamically adjust traffic signal timing. Potential improvements include more advanced algorithms and incorporating connected- and automated-vehicle data. There is an opportunity to use real-time analytics to improve dynamic, adaptive signal timing.

5 RESEARCH CHALLENGES

In the previous sections we described transportation data users and potential applications. In this section, we describe research challenges in building a transportation-data management system. We argue that there are challenges across the entire stack, including data model, data integration, physical data stores, query processing, data dynamics, data re-purposing, and standardization.

5.1 Unified Data Model

In recent years, graph-based data models have been used in big-data problems where the data could be naturally mapped to entities and relationships between those entities. Indeed, the transportation data from different sources has an underlying transportation network imbued with multiple relationships types, e.g. “Rita *likes* Dutch Brothers Coffee Shop and *knows* Hector". Analysts want to inherently leverage that network and co-explore different types of relations in their analysis. However, such co-explorations lead to expensive joins based on foreign key relationships in the traditional relational model. The idea of using graphs is to walk across the data structure, i.e., do look-ups instead of performing joins.

A graph is a collection of nodes or vertices and links or edges between them, which may have one or more meta-data attributes. For example, in transportation, bus schedules could be directly mapped to a graph with bus stops as vertices and bus routes between stops as edges. Such a graph may have additional meta-data indicating whether the stops (vertices) have shelters or not and whether the bus connections (edges) have bicycle racks or not. Other data sources might need to be preprocessed into graphs. For instance, we can aggregate the average speed recorded by speed sensors on different road segments and represent them as edges on the road network. This preprocessing step embeds the large join or group by operation required to establish the links between two objects, which is analogous to establishing the foreign key relationships. However, the cost is justified as it occurs one time during ingestion [27] as opposed to repeatedly doing joins during query processing in the relational model. Also, the dynamic merging of multiple geo-spatial data sets is a very difficult problem. Systems such as OpenLR show promise in addressing the merging issues [36].

Analysts need to view data from different sources as a single *connected multi-graph*, which essentially means that we connect (logically) different types of entities and their relationships from individual graphs into a single unified graph. Here the join operator to unify the graph could be a straightforward equality match, e.g., joining train and bus networks on their common stops (vertices). Alternatively, we could also apply more fuzzy join conditions, e.g., joining coffee and bus networks based on the geographic proximity of their vertices. The multi-graph data model can provide a unified view of the transportation data while letting the system take care of the integration.

5.2 Just-In-Time Data Integration

Typically, transportation data collection and use has been siloed by transportation agencies. However, advances in technology and more sophisticated management of transportation systems are beginning to require real-time, cross-agency integration of this data. The Multi-modal Routing, Improved Transit Arrival Predictions and Integrated Corridor Management applications from the previous section require cleaning and integrating large volumes of data of different types and from different sources in real time.

Data integration has been long studied in computer science [4, 5, 31]. Traditional data integration aims to integrate several source schemas into a single final schema. However, transportation data sources cannot be converted to a new "final" schema; the data sources will continue to produce data in their respective schemas — and continue to evolve — so an on-line integration approach is needed. Furthermore, different users have different integration requirements. For example, transportation planners want to analyze traffic congestion data for all freeways in a city sorted over time and grouped by month. This sorting operation is starkly different from system operators who want to view operational data from all buses around an intersection where an accident occurred a few seconds ago. It may be that every user — or even every task — needs data integrated in a different way.

Thus, there is a need for a *just-in-time* approach to data integration, that considers both the real-time data as well as the real-time user requirements. For example, an Integrated Corridor Management (ICM) application may need to adjust bus dispatch or re-time traffic signals on an arterial to accommodate an incident on the freeway. ICM must take the input data from transit, DOT and city and county agencies and produce output data customized for each of the three agencies. Recent efforts, such as lazy ETL [28], do lower the data load and integration costs; however, they would need to be adapted to an on-line setting for continuously arriving data.

5.3 Choosing an Efficient Storage System

Transportation data sets are highly heterogeneous — with operational (i.e. transactional), streaming and archived data. The design decision to select an efficient storage subsystem to manage such diverse data or whether to create a new system is an open research question. The traditional *no one-size-fits-all* [46] approach would suggest using separate database backends to manage the diverse data. In this direction, two alternate approaches have been proposed in the recent past. The first is to build one-size-fits-all store that is self-adaptive and automatically configures the backend store for the current workload at hand. Examples include OctopusDB [8] and H2O [1]. The second approach is to build middle-ware that efficiently combines multiple backend stores into one seamless data view for the user. Examples of this include the BigDawg Polystore [11, 12], and invisble glue [2].

Applying the approaches above to transportation data, one could create one or more physical views to map the transportation multi-graph data to storage back-ends leading to better query performance. For example, a multi-graph can be stored in part as relational tables, multidimensional arrays, de-normalized flat files, standardized ontologies using Resource Description Framework (RDF) triplets [21], or simply as key-value pairs. In addition, approaches such as creating secondary physical views on the same data, i.e., the same piece of data represented in multiple ways physically by creating secondary indexes and other materialized views can also be pursued. Selecting the right physical views (both primary and secondary) is challenging because of the large design space (there are multiple data sources which could be combined and stored in a large number of ways), lack of a fixed query workloads (transportation workloads are often exploratory or ad-hoc in nature), and the presence of multiple storage backends. Exploring these different approaches and picking the best storage back ends for transportation data is an interesting research challenge.

5.4 Multi-graph Query Processing

Processing queries over the transportation data introduces a number of challenges. First, traditional graph analytics involves loading and analyzing a given (typically static) graph. In contrast, as described earlier, the transportation dataset is essentially a graph of graphs and one or more graphs may be dynamically loaded and processed depending on the analysis. For example, an analyst may look for routes where every transfer has a coffee shop, thereby involving the bus and the coffee shop graph. This ad-hoc selecting and combining different pieces of the multi-graph is challenging.

Second, as a result of the complexity of the data sources, processing descriptive declarative queries on transportation data is challenging. For instance, a user may need to quickly find a route rather than waiting minutes to find the best route. Thus, there is a trade-off between producing fast answers versus good answers. Furthermore, the complexity of the queries mean that the best or exact answer may not be always possible, e.g., no route may satisfy all conditions in case of personalized routing. In such situations, the system could consider producing non-empty results which satisfy the maximum number of predicates, ask users to rank the predicates, or even decorate predicates as (soft) preferences or (hard) constraints. These design choices need to be explored in more detail.

Third, query processing on multi-graphs blows up the search space for many of the graph algorithms. For example, path finding algorithms now need to consider several path combinations between each of the input graphs. This makes finding the right answers difficult for the users. In many cases, a user may want to try out several sets of predicates before settling on ones which best satisfy his requirements and which could be computed in a reasonable time. The question then is whether the system can help users to iterative explore the search space, e.g., slice and dice it into regions of interest, and discover their results [44].

Finally, transportation data analytics could involve multiple workload types over the same data set. Examples include supporting both real-time (for monitoring and trouble shooting) as well as off-line (for reporting and planning) analytics, supporting graph queries (e.g. identifying congestion) and relational operations (e.g. joining two or more graphs) on the same underlying data, and running both batch as well as streaming analytics. Supporting these mixed workload scenarios is challenging and there are several recent efforts to build multi-workload and multi-store systems. Applying those ideas to the transportation domain, where the data comes from a variety of data sources and maps naturally to a graph structure, is an interesting problem.

5.5 Data Dynamics

Any transportation data system is sensitive to time as parts of the data are frequently updated, such as traffic, load, etc. The queries have a time dependency and a predictive nature as well. For example, a user may want to wait for bus on which he or she can get a seat, but will the bus be still empty by the time the user actually boards it? What will be the traffic conditions be by the time a user reaches the road segment? And so on.

Thus, transportation data systems needs to (i) efficiently manage updates to portions of the multi-graph, (ii) consider the timing of the transportation activity and how the data might change in the near future, and (iii) consider the tolerance for stale results for different users/applications, i.e., what should be the refresh rate in case new data arrives? Note that different data sources have different update schedules. Data streams from on-road or vehicular sensors are updated multiple times a second or minute, crash information may be on line for larger crashes, but may be delayed for smaller incidents that do not have a major impact on traffic. Transit schedules may change quarterly or more frequently, e.g. in Portland bus schedules are refreshed every two weeks. Some data has reporting schedules, for example, data from some sensors might be reported every 6 hours, whereas events such as crashes or inclement weather come up as they occur. While a traditional database system can clearly handle infrequent updates to static data such as transit schedules and sensor locations, a specialized time series database e.g. IBM Informix are required to efficiently order and index time series data availed from the sensors. It is an additional challenge to combine static data with the streamed data.

In addition to automated responses to time-sensitive data by the transportation data system, the users may want to influence some of these decisions. For example, Starbucks may want a new outlet to be immediately discoverable. Other updates can be driven by the user, e.g., a user may want to try any new coffee shop on the way, thereby requiring the system to update the coffee shop graph. We believe that these requirements need an efficient meta-data management system and standardized interfaces to represent the data changes from these different organizations.

5.6 Data Re-Purposing

As described earlier, transportation data is most often collected for the purpose of operating the transportation system and this "operational data" is often archived and then used for planning, analysis, and research; we call this use of data collected for one purpose and used for another, referred to as *data re-purposing*. A primary motivation for data re-purposing is opportunistic — that is to take advantage of readily-available, low-cost data with good temporal and spatial coverage. Operational data has very broad coverage compared to what is typically available for planning, analysis and research. For example, a targeted traffic study completed as part of a planning or research process might collect 48 hours of counts at a particular location; in contrast, operational data gives 24x7x365 data at hundreds of locations across a city.

While re-purposing data allows planners and analysts to use low-cost, high-coverage data sources, data re-purposing is challenging. Specifically re-purposed data typically does not exactly match the user needs. For example, data collected for the purpose of displaying travel times to the general public might be re-purposed and used in a research study to identify locations of traffic bottlenecks. As such, the locations of the sensors (which were placed for the purpose of travel time calculation), will not be ideal for the bottleneck study. Similarly, the data quality may be lower than ideal for the bottleneck study. However, due to coverage and low cost, the operational data is still valuable for the bottleneck research study. Techniques for assisting researchers and planners in bridging the gap between the needs of their study and the available, but not perfect, data are needed. Lastly, institutional barriers regarding the sharing of data may prevail in some areas.

5.7 Standardization & Architectures

The standardization of transit data through the development of the GTFS format [19] enabled the integration of transit routing information in the Google Maps and other apps. In current practice, transportation data is collected from a variety of sources. Data is collected from multiple agencies, but more importantly different agencies use different providers for the same service. For example, the PORTAL archive receives traffic signal data from three different vendors. The data produced by all three vendors is similar in content, but different in structure and in details. These differences lead to inefficiency during combination. Standards or specifications must be developed for all types of transportation data to enable its effective usage. As demonstrated by TriMet's (the Portland, OR regional transit agency) experience with GTFS, successful standardization efforts require community engagement and simplicity.

6 PORTLAND URBAN DATA LAKE (PUDL)

The Portland Urban Data Lake (PUDL) is a collaborative project which aims to begin to address some of the challenges described above. The goal of the PUDL project is to develop Urban Analytics for use by decision makers and to provide data access to public agencies and the general public, especially innovators in the tech community. From a policy perspective, PUDL will contribute to increased transparency through open data and improved safety and improved mobility through urban analytics, especially for traditionally underserved communities. From a technical perspective, this project will collect, store and integrate Smart Cities-related data and other data from a variety of sources including new sensor deployments and existing data sources to provide a foundation for data-driven decision making in the City of Portland. The project is a collaboration between the City of Portland, Portland State University, TriMet, Metro and Portland General Electric.

A key goal of PUDL is to develop a data architecture and platform that can integrate and fuse the many Smart City data sources that

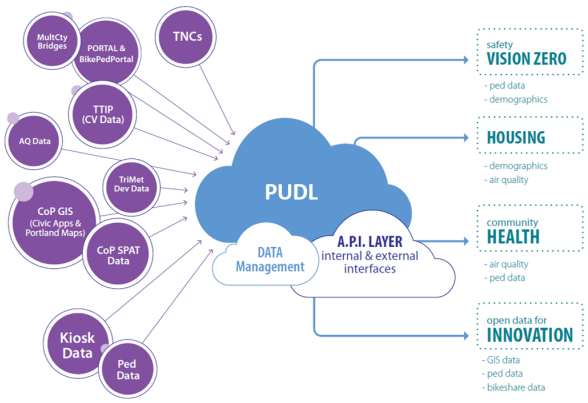

Figure 2: Proposed PUDL Architecture

are available and that are required to effectively develop Urban Analytics to achieve effective data-driven decision making. PUDL aims to use the new architecture and platform to enable the effective use of Data. At this time, PUDL pilot projects are underway with community-based organizations and private sector partners. Figure 2 shows the proposed PUDL architecture.

7 CONCLUSION

The transportation domain presents significant challenges and opportunities for data management researchers. The vast store of transportation data has yet to be fully leveraged and presents a variety of interesting research areas with potentially large impacts. We have presented a description of the transportation data domain, personas that may use the transportation data along with their potential applications, and a detailed research landscape for building a transportation data system. Open problems for the data management community include unified data model, just-in-time data integration, choosing efficient storage systems, multi-graph query processing, handling dynamic transportation data, and data re-purposing.

8 ACKNOWLEDGMENTS

The authors acknowledge the Intel Science and Technology Center for Big Data (ISTC-BD) and the Portland State University Institute for Sustainability Solutions (ISS) for their support of this work.

REFERENCES

[1] Ioannis Alagiannis, Stratos Idreos, and Anastasia Ailamaki. 2014. H2O: A Hands-free Adaptive Store. In *SIGMOD*. 1103–1114.
[2] Francesca Bugiotti, Damian Bursztyn, Alin Deutsch, Ioana Ileana, and Ioana Manolescu. 2015. Invisible Glue: Scalable Self-Tuning Multi-Stores. In *CIDR*.
[3] PDX Bus. 2016. PDX Bus. http://pdxbus.teleportaloo.org. (2016).
[4] CL Philip Chen and Chun-Yang Zhang. 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences* 275 (2014), 314–347.
[5] Jinchuan Chen, Yueguo Chen, Xiaoyong Du, Cuiping Li, Jiaheng Lu, Suyun Zhao, and Xuan Zhou. 2013. Big data challenge: a data management perspective. *Frontiers of Computer Science* 7, 2 (2013), 157–164.
[6] Atlanta Regional Commission. 2015. Atlanta Regional Commission. http://www.atlantaregional.com/plan2040. (2015).
[7] Brian Cronin. 2012. Vehicle Based Data and Availability. http://www.its.dot.gov/itspac/october2012/PDF/data_availability.pdf. (2012).
[8] Jens Dittrich and Alekh Jindal. 2011. Towards a One Size Fits All Database Architecture. In *CIDR*. 195–198.
[9] Google[x] Dr. Chris Urmson; Director, Self-Driving Cars. 2015. 2015 ITS America Annual Meeting Opening Plenary. https://www.youtube.com/watch?v=9m0xMeWONhs. (June 2015).
[10] DriveNET. 2015. DriveNET. http://www.uwdrive.net. (2015).
[11] Jennie Duggan, Aaron J. Elmore, Michael Stonebraker, Magda Balazinska, Bill Howe, Jeremy Kepner, Sam Madden, David Maier, Tim Mattson, and Stan Zdonik. 2015. The BigDAWG Polystore System. *SIGMOD Rec.* 44, 2 (2015), 11–16.
[12] Aaron Elmore et al. 2015. A Demonstration of the BigDAWG Polystore System. *PVLDB* 8, 12 (2015).
[13] EPA. 2017. EPA Climate Change. https://19january2017snapshot.epa.gov/ghgemissions/sources-greenhouse-gas-emissions_.html. (2017).
[14] HOP Fastpass. 2015. hop fastpass. http://myhopcard.com. (2015).
[15] L. Gasparini, E. Bouillet, F. Calabrese, O. Verscheure, B. O'Brien, and M. O'Donnell. 2011. System and analytics for continuously assessing transport systems from sparse and noisy observations: Case study in Dublin. In *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on*. 1827–1832. https://doi.org/10.1109/ITSC.2011.6082982
[16] Google. 2015. The bright side of sitting in traffic: Crowdsourcing road congestion data. https://googleblog.blogspot.com/2009/08/bright-side-of-sitting-in-traffic.html. (2015).
[17] Google. 2016. Google. http://www.google.com. (2016).
[18] Google. 2016. Google Self-Driving Car Project. https://www.google.com/selfdrivingcar/. (2016).
[19] GTFS. 2016. General Transit Feed Specification. https://developers.google.com/transit/gtfs/. (2016).
[20] HERE. 2018. here. https://www.here.com/en. (2018).
[21] Jiewen Huang, Daniel J Abadi, and Kun Ren. 2011. Scalable SPARQL querying of large RDF graphs. *Proceedings of the VLDB Endowment* 4, 11 (2011), 1123–1134.
[22] ITS Public Data Hub. 2018. ITS Public Data Hub; Federal Highway Administration. https://www.its.dot.gov/data/. (2018).
[23] Inrix. 2018. Inrix. http://inrix.com/. (2018).
[24] iPeMS. 2015. iPeMS. http://www.iteris.com/products/software/iterispems-ipems. (2015). Accessed: Oct-12-2015.
[25] ITS JPO, USDOT. 2016. ITS Research 2015-2019: Connected Vehicles. http://www.its.dot.gov/research_areas/connected_vehicle.htm. (2016).
[26] Nikola Ivanov. 2016. Big Data in Transportation. http://itsmd.org/wp-content/uploads/Nikola-Ivanov-Big-Data-in-Transportation.pdf. (2016).
[27] Nilesh Jain, Guangdeng Liao, and Theodore L. Willke. 2013. GraphBuilder: Scalable Graph ETL Framework. In *First International Workshop on Graph Data Management Experiences and Systems (GRADES '13)*. ACM, New York, NY, USA, Article 4, 6 pages. https://doi.org/10.1145/2484425.2484429
[28] Yağız Kargın, Milena Ivanova, Ying Zhang, Stefan Manegold, and Martin Kersten. 2013. Lazy ETL in Action: ETL Technology Dates Scientific Data. *PVLDB* 16 (2013).
[29] Kartik Kaushik. 2014. Using NPMRDS: Lessons Learned. http://nationalruralitsconference.org/downloads/Presentations14/Kaushik_A3.pdf. In *National Rural ITS Conference, 2014*. ITS America. Accessed: Oct-12-2015.
[30] Sidewalk Labs. 2016. Sidewalk Labs. http://www.sidewalklabs.com. (2016).
[31] Steve LaValle, Eric Lesser, Rebecca Shockley, Michael S Hopkins, and Nina Kruschwitz. 2013. Big data, analytics and the path from insights to value. *MIT sloan management review* 21 (2013).
[32] Oregon Green Light. 2015. Oregon Green Light. http://www.oregon.gov/odot/mct/pages/green.aspx. (2015).
[33] Lyft. 2016. Lyft. http://www.lyft.com. (2016).
[34] NHTSA. 2016. 2016 Fatal Motor Vehicle Crashes: Overview. https://crashstats.nhtsa.dot.gov/Api/Public/Publication/812456. (2016).
[35] NPMRDS. 2015. NPMRDS. http://www.ops.fhwa.dot.gov/freight/freight_analysis/perform_meas/vpds/npmrdsfaqs.htm. (2015).
[36] OpenLR. 2016. OpenLR. http://www.openlr.org. (2016).
[37] ORCA. 2015. ORCA Card. https://orcacard.com/. (2015).
[38] PORTAL. 2015. Multi-modal data set for Portland Oregon Region Test Data Set for the FHWA Connected Vehicle Initiative Real-Time Data Capture and Management Program. http://portal.its.pdx.edu/Portal/index.php/fhwa. (2015).
[39] PORTAL. 2015. Portal Transportation Data Archive. http://portal.its.pdx.edu. (2015).
[40] RideScout. 2015. RideScout. http://www.ridescoutapp.com. (2015).
[41] RITIS. 2015. RITIS. http://www.cattlab.umd.edu/?portfolio=ritis. (2015).
[42] SAE. 2015. Dedicated Short Range Communications (DSRC) Message Set Dictionary Support Page. http://www.sae.org/standardsdev/dsrc/. (2015).
[43] SCATS. 2015. SCATS. http://www.scats.com.au. (2015).
[44] Thibault Sellam and Martin Kersten. 2013. Meet Charles, big data query advisor. In *CIDR*.
[45] Sidewalk. 2018. Sidewalk Labs. https://www.sidewalklabs.com/. (2018).
[46] M. Stonebraker and U. Cetintemel. 2005. "One size fits all": an idea whose time has come and gone. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*. 2–11. https://doi.org/10.1109/ICDE.2005.1
[47] Streetlight. 2018. StreetLightData. https://www.streetlightdata.com/. (2018).
[48] Tesla. 2016. Tesla Motors Autopilot. https://www.teslamotors.com/presskit/autopilot. (2016).
[49] TriMet. 2018. TriMet. trimet.org. (2018).
[50] TTI. 2015. TTI 2015 Urban Mobility Scorecard. (2015).
[51] Kristin Tufte, Robert Bertini, and Morgan Harvey. 2015. Evolution and Usage of the Portal Data Archive: A Ten-Year Retrospective. *Transportation Research Record: Journal of the Transportation Research Board* (2015).
[52] Uber. 2016. Uber. http://www.uber.com. (2016).
[53] USDOT. 2014. Overview of the USDOT Real-Time Data Capture and Management (DCM) Program. http://www.itsa.wikispaces.net/file/view/1.+CV+and+DCM+overview.pdf. (2014).
[54] USDOT. 2017. USDOT Releases 2016 Fatal Traffic Crash Data. https://www.nhtsa.gov/press-releases/usdot-releases-2016-fatal-traffic-crash-data. (2017).
[55] USDOT - ITS JPO. 2015. Connected Vehicle Frequently Asked Questions. http://www.its.dot.gov/connected_vehicle/connected_vehicles_FAQs.htm. (2015).
[56] Waze. 2015. Waze. https://www.waze.com. (2015).